Heart Disease Analysis and Prediction

Multivariate Analysis Project

Wei, Chia-Yu

1 Introduction

The term “heart disease” includes several diseases about heart condition, such as heart failure, heart attack, irregular heartbeats, or some problem about heart muscle. According to the CDC, heart disease is the leading cause of death across gender and racial/ethnic groups in United State, with one person dying every 33 seconds from cardiovascular disease. Heart disease can have significant impacts on only family but country. Early diagnosis and treatment of heart disease can significantly reduce the occurrence and severity of the disease. The question is “can we predict if the patients will suffer from heart disease through some features”.

1.1 Research Questions

- Figure out the factor of heart disease.
- Analyze the relation between variables.
- Predict the heart disease status by the attributes.

1.2 Data Information

The data was collected by 918 patients. There are 9 variables. Here is the description:

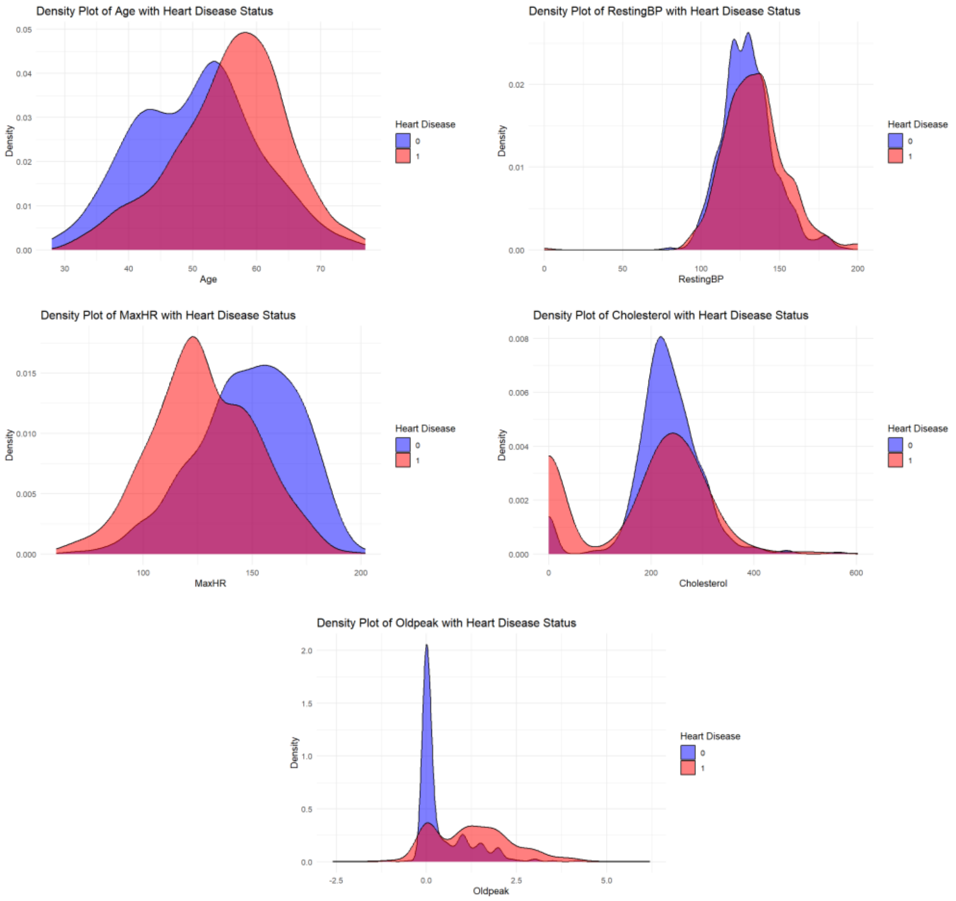
- X1 Age
- X2 Sex (1: Male; 0: Female)
- X3 RestingBP: resting blood pressure
- X4 Cholesterol
- X5 FastingBS: fasting blood sugar
- X6 MaxHR: maximum heart rate achieved
- X7 ST depression induced by exercise relative to rest
- X8 ExerciseAngina: exercise induced angina (1: yes; 0: no)
- X9 ST_slpoe: the slope of the peak exercise ST segment
  - 0: Upsloping
  - 1: Flat
  - 2: Downsloping
- X10 HeartDisease (presence of heart disease)

1.3 Data Overview

1.2.1 Continuous variables

I draw the destiny for all numeric variables. The points to notice are that the heart disease is more prevalent in older age group. Moreover, the patients with lower heart rate are more likely to be detected having heart disease. We can see that there are so many zero in the Cholesterol, which is not normal situation. Therefore, I decided to remove these observations.

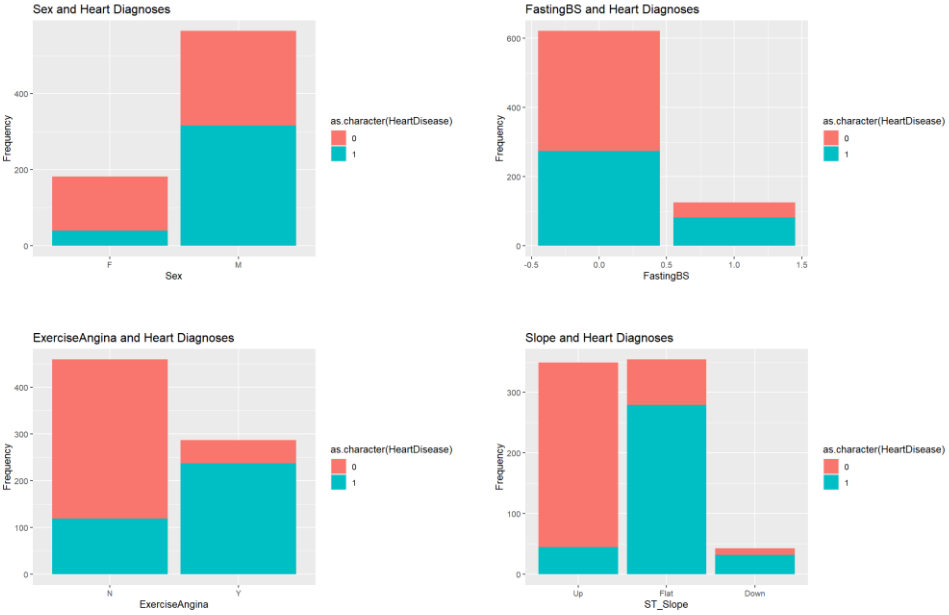
Figure 1: Density Plot of Numeric Variables



1.2.2 Discrete Variables

Based on the plots I obtained (Figure 2), the patient who has exercise-induced angina has higher probability to get heart disease. Also, if the patients' ST_slope is flat or downsloping, they are likely to have heart disease.

Figure 2: Plots for Categorical Variables



2 Logistic Regression

The linear model is used when the response is continuous. However, the response I chose is binomial, there are two possible outcomes with probability distribution.

$$P(Y = 1) = p$$
$$P(Y = 0) = 1 - p$$

Instead of fitting a line to the data, logistic regression fits logistic function from 0 to 1 and that means the probability that the patient gets heart disease.

2.1 Regression Equation

The regression model estimates the probability of $Y = 1$. The logistic regression is like linear regression can work with continuous data and discreate data. Therefore, based on my data, we have the model

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1(Age) + \beta_2(SexM) + \cdots + \beta_7(ST_{SlopeFlat}) + \beta_8(ST_{SlopeDown})$$

Equivalent,

$$P(Y = 1) = \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}}$$

2.2 Regression Equation

I fit the initial logistic model for probability of getting heart disease:

$$\log\left(\frac{p}{1-p}\right) = -5.534 + 0.031(Age) + 1.847(SexM) + 0.0087(RestingBp)$$

$$+ 0.0036(Cholesterol) + 0.2233(FastingBS) - 0.0.57(MaxHR)$$

$$+ 0.382(ExerciseAnginaY) + 0.45(Oldpeak) + 1.3(SlopeFlat)$$

$$+1.18(SlopeDown)$$

- AIC: 549.17

Based on the model I got, all the variables have positive effect on the presence of heart disease except maximum heart rate achieved. The meaning of estimated coefficient is that how the probability of having heart disease compare to no disease changes when the variable increase one unit. Take Sex for example, if the patient is male, the log ratio of having heart disease increases by 1.1847 with other variables unchanged.

2.3 Step-wise Regression

Step-wise regression is a popular method for subset selection. I decided to use stepwise selection to see what variables are significant to the response variable.

$$\log\left(\frac{p}{1-p}\right) = -6.9722 + 0.04(Age) + 1.8818(SexM) + 0.0038(Cholesterol)$$

$$+ 1.4599(ExerciseAnginaY) + 0.4555(Oldpeak) + 2.532(SlopeFlat)$$

$$+ 1.2482(SlopeDown)$$

- AIC: 546.81

2.4 Diagnostics

There may some problems such as multicollinearity, outliers, influential observations etc. Further, there are some assumptions we have to check.

2.4.1 Multicollinearity

To check multicollinearity, I obtained variance inflation factor (VIF). The values of each variable are very close to 1. It indicates that the model doesn’t have multicollinearity.

Variables	X1	X2	X4	X5	X6	X7
VIF	1.1316	1.0549	1.1353	1.0768	1.1897	1.0974

Table 1: Variance Inflation Factors

2.5 Accuracy and Sensitivity

I divided the data into two groups, train and test data, then classified predicted probabilities greater than 0.5 as class 0 (no heart disease). Next, I created a confusion matrix, which is a table of true value vs. predicted values.

	0	1
0	59	9
1	10	71

Table 2: Confusion Matrix

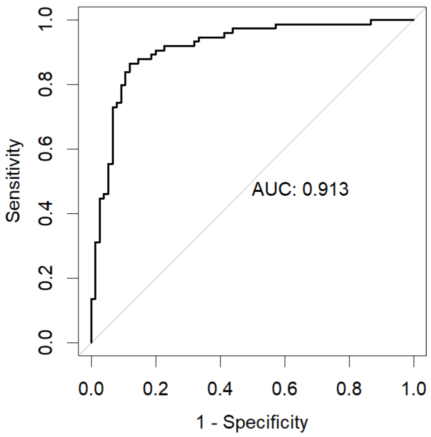


Figure 3: ROC Curve Plot

- Accuracy: 0.87248
- AUC: 0.913

3 Principal Component Analysis

3.1 Preparing Data for Principal Component Analysis

Before Applying principal component analysis to my data set, I preformed standardization for numeric variables. Then I obtained covariance matrix.

Age	RestingBP	Cholesterol	MaxHR	Oldpeak
-1.36	0.41	0.75	1.3	-0.84
-0.41	1.56	-1.09	0.64	0.09
-0.67	-0.17	0.64	-1.72	-0.84
-0.51	0.28	-0.52	-1.31	0.56
0.12	0.98	-0.84	-0.74	-0.84

Table 3: Standardized values of first five observations of Variables

3.2 Proportion and Cumulative Proportional Variance of PCs

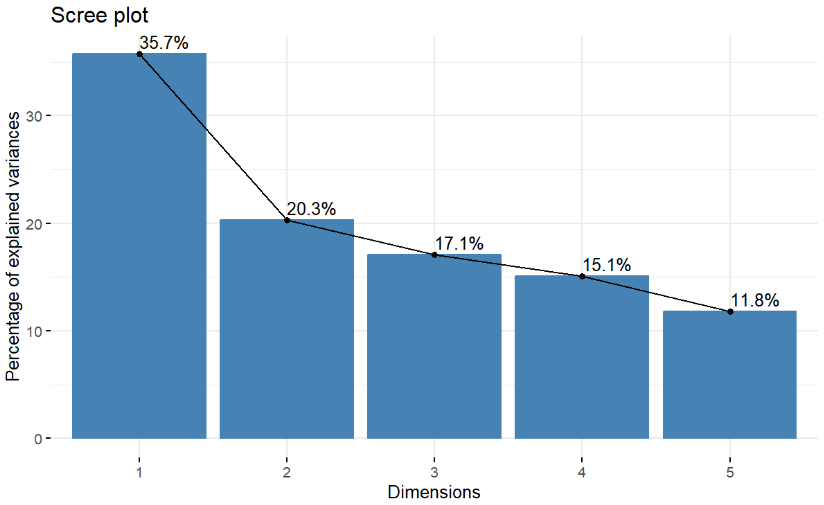
I obtained the proportion of variance and cumulative proportional variance of each principal component. The 88% of variance can be explained by first four principal components.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.337	1.007	0.924	0.869	0.768
Proportion of Variance	0.357	0.203	0.171	0.151	0.118
Cumulative Proportion	0.357	0.56	0.731	0.882	1

Table 4: Variance Contribution of Principal Component

3.3 Scree Plot

Figure 4: Scree Plot



3.4 Correlation between Standardized Variables and Principal Components

The values in matrix represents a principal component (PC), and each row corresponds to one of the original variables. The values in the matrix indicate how much each original variable contributes to each principal component. Take first component for example, all numeric variables contribute positive effect to PC1 except MaxHR.

	PC1	PC2	PC3	PC4	PC5
Age	0.567	-0.119	-0.057	0.331	0.742
RestingBP	0.408	0.306	0.79	0.184	-0.284
Cholesterol	0.141	0.891	-0.429	0.032	-0.011
MaxHR	-.0505	0.307	0.432	-0.31	0.607
Oldpeak	0.486	-0.058	-0.024	-0.871	0.005

Table 5: Correlation between variables and PCs

4 Canonical Correlation Analysis

I applied canonical correlation analysis to see the correlation between a combination of variables and another combination of variables. In this analysis, I group the variables into three groups.

4.1 Variable Groups

I grouped the variables by their information. There are three groups and heart disease.

- Demographic Factors
  - Age
  - Sex
- Heart Function
  - Oldpeak
  - ExerciseAngina
  - ST_Slope
- Resting Physiological Measures
  - Cholesterol
  - RestingBP
  - FastingBS
  - MaxHR
- Heart Disease

4.2 Canonical Correlation

From the Table 5, for resting physiological measures, 49% of variance is explained by the demographic variables; 42% is explained by heart disease and only 23% is explained by variables about heart function.

	Demographic	Heart Function	Physiological Measures	Heart Disease
Demographic Factors	1	0.37	0.49	0.41
Heart Function	0.37	1	0.23	0.67
Physiological Measures	0.49	0.23	1	0.42
Heart Disease	0.41	0.67	0.42	1

Table 6: Highest canonical correlation between Groups

4.3 Discussion

The values in the tables represent the correlation between different groups of variables. As we can see, all groups are highly correlated with heart disease. The variables about heart function affect the presence of heart disease a most.

5 Cluster Analysis

Clustering is a method to group data into sets so that the elements in same group are more similar to each other. Then we apply these methods to determine what groups the new data is in. I used two different algorithms:

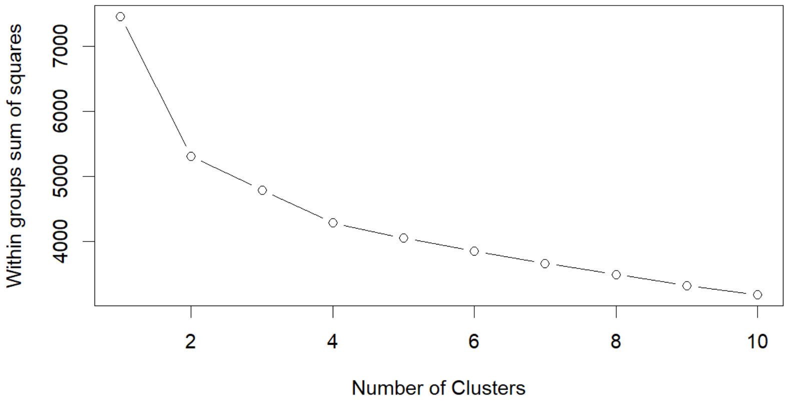
- K-means
- Support Vector Machine

5.1 K-means

5.1.1 Elbow Method

I used elbow method to determine the number of clusters. It is a common method to be used before doing k-means. We selected the number of cluster where elbow seen. Based on the Figure 5, I chose k = 2.

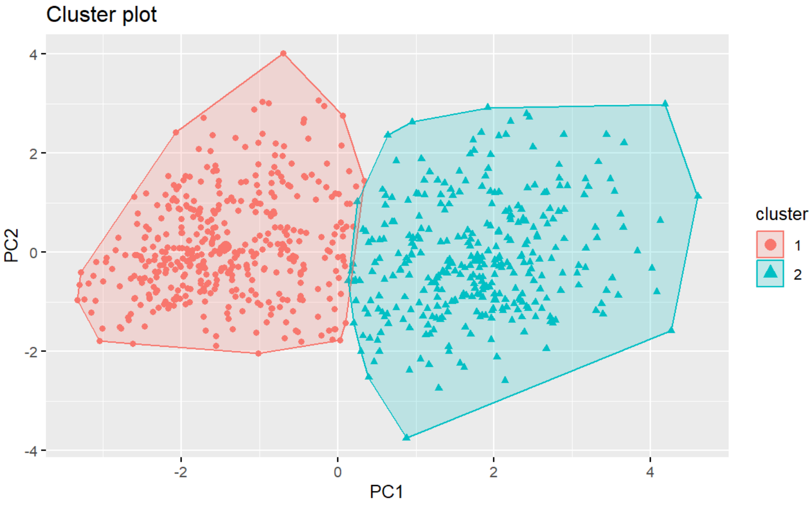
Figure 5: Elbow Plot



5.1.2 Results

I did k-means with k=2 since there are two possible outcome (have/no heart disease).

Figure 6: K-means (k=2)



- Accuracy: 0.9035

5.2 Support Vector Machine

I trained svm model, then use trained model to make prediction on the test data set. Also, I provided the confusion matrix of true versus predicted lables:

	0	1
0	60	8
1	12	69

Table 7: Confusion Matrix

- Accuracy: 0.8658

5.3 Comparison

Compare the accuracy of three methods:

Logistic regression	0.8725
K-means	0.9035
Support Vector Machine	0.8658

Table 8: Accuracy from three models

From the result we get from the parts above, the k-means seems that best model for the data set.

5.4 Discussion

I applied two clustering algorithms: K-means and SVM, and compare them with the result from logistic model. All accuracies of them are good and similar. My thinking is that we can do further research with larger data set since the data has only 746 observations. I believe that we can see some difference to find the best model.

6 Conclusion

In the project, we explore some likely factor of heart disease first and built the logistic model to calculate the probability of having heart disease. Then applied some clustering algorithms to predict the heart disease status based on the attributes.

I we separately looked at the relation between each variable and response variables (presence of heart disease) first. From the plots we got, the patients with older age or lower heart rate are more likely to be detected to have heart disease. Moreover, if the patients have angina during exercise or their ST segment are flat or downsloping, they have high probability of getting disease.

Next, we create the logistic model which tells us that all the variables were kept in the final model have positive effect for response variables. In PCA and CCA parts, we see how numeric variables effect the response and how a set of variables effect another set of variables. It seems that the variables about heart function are highly correlated with having the heart disease or not. Also, the different sets affect each other interactively.

Last, we applied k-means and support vector machine algorithm to see which method is better for predicting the disease status. The k-means seems the best way. However, I think we should apply these methods with larger data set in the future to make the more precise conclusion.

References

Heart Failure Prediction [Dataset]
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Jen Thomas. Heart Disease: Fact, Statistics, and You, 2023
https://www.healthline.com/health/heart-disease/statistics#What-are-the-risk-factors?

Anish Singh Walia. Radial kernel Support Vector Classifier, 2017
https://datascienceplus.com/radial-kernel-support-vector-classifier/